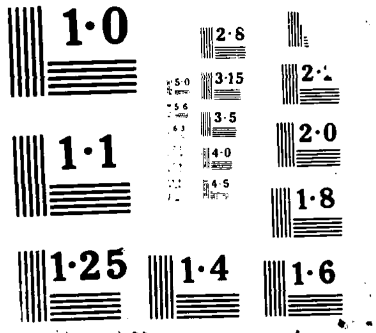
AD-A193 189 A NOTE ON PROFILE LIKELIHOOD LEAST FAVOURABLE FAMILIES 1/1
AND KULLBACK-LEIBL (U) STANFORD UNIV CA DEPT OF
STATISTICS R TIBSHIRANI ET AL 19 APR 88 TR-404

UNCLASSIFIED N00014-86-K-0156 F/G 12/3 NL


1·0
2·8
3·15
3·5
4·0
4·5
2·0
1·1
1·8
1·25
1·4
1·6

A NOTE ON PROFILE LIKELIHOOD, LEAST FAVOURABLE FAMILIES AND KULLBACK-LEIBLER DISTANCE

BY

ROBERT TIBSHIRANI and LARRY WASSERMAN

TECHNICAL REPORT NO. 404
APRIL 19, 1988

PREPARED UNDER CONTRACT
N00014-86-K-0156 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

4

A NOTE ON PROFILE LIKELIHOOD, LEAST FAVOURABLE FAMILIES AND KULLBACK-LEIBLER DISTANCE

BY

ROBERT TIBSHIRANI and LARRY WASSERMAN

TECHNICAL REPORT NO. 404
APRIL 19, 1988

Prepared Under Contract
N00014-86-K-0156 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A NOTE ON PROFILE LIKELIHOOD, LEAST FAVOURABLE FAMILIES AND KULLBACK-LEIBLER DISTANCE

Robert Tibshirani and Larry Wasserman

SUMMARY

We consider several methods for reducing high dimensional models to one dimensional models for the purpose of simplifying likelihood inferences. The equivalence between these methods is investigated.

Some Key Words : nuisance parameters, likelihood, exponential families.

1. INTRODUCTION

Consider a statistical model Γ consisting of a class of densities $\{f(x \mid \eta)\}$ where $\eta \in \Omega \subset R^k$ is a vector-valued parameter of dimension greater than one. Often we are interested in a real valued function $\theta \equiv \theta(\eta)$. Many useful inferential techniques involve the log-likelihood function defined by

$$L(\eta) = a + \sum_i \log(f(x_i \mid \eta))$$

where $x_1, x_2, \cdots, x_n$ are independent observations from the true density and a is an arbitrary real constant which, for convenience we shall take to be zero. In general, a one-

dimensional likelihood function is not available for the parameter of interest θ. The problem of constructing such a likelihood function is discussed at length by Kalbfleisch and Sprott (1970).

The method that we discuss here for dealing with this problem is to strategically choose a sub-family of densities from Γ indexed by θ. We then construct a likelihood function based on the new reduced model. Specifically, let $\Gamma_\theta = \{f(x \mid \theta)\}$ denote the reduced model. (When convenient, Γ_θ will also refer to the corresponding curve in the parameter space Ω). We then take $L(\theta) = \sum \log(f(x_i \mid \theta))$. (Unqualified sums are to be taken from i=1 to n).

We shall consider several such techniques for choosing Γ_θ and investigate certain equivalences between them. We note that some of the methods of model reduction that we discuss were originally proposed for reasons other than constructing likelihood functions.

2. DEFINITIONS

The first reduction technique we consider is used to construct the profile likelihood (see Kalbfleisch and Sprott, 1970). Let $f(x \mid \theta)$ be the density which maximizes the probability of the observed data subject to $\theta(\eta) = \theta$. This defines a family indexed by θ which we will denote by Γ_θ^{PL}. The resulting log-likelihood function will be denoted by $L^{PL}(\theta)$. Note that $L^{PL}(\theta)$ passes through the global maximum of $L(\eta)$.

Another method of defining a one parameter family is what Stein (1956) calls the "least favourable family" given by

$$\eta(\tau) = \hat{\eta} + \tau\, I_\eta^{-1} \nabla\theta(\hat{\eta})$$

where $I_{\hat\eta}$ is the observed Fisher information at $\hat\eta$ and $\nabla\theta(\eta)$ is the vector whose i^{th} component is $\partial\theta/\partial\eta_i$. This traces a straight line through $\hat\eta$ having direction $I_{\hat\eta}^{-1}\nabla\theta(\eta)$. Denote this family by Γ_θ^S and the corresponding log-likelihood by $L^S(\theta)$. This family has the property that the observed Fisher information for θ is the same as in the full family $f(x\,|\,\eta)$ (Stein, 1956). Furthermore, any other (linear) sub-family through $\hat\eta$ has greater Fisher information for θ. (Note that Stein used expected information in his definition. Here we follow Efron (1984) and use the observed information).

Still another reduction method is employed by Efron (1981,1984) for the purpose of constructing confidence intervals in multi-parameter and non-parametric settings. Let

$$C_{\theta_0} = \{\eta : \theta(\eta) = \theta_0\},$$

the level surface of constant θ. Efron selects the value of η from C_{θ_0} such that the Kullback-Leibler distance

$$K(\hat\eta,\eta) = \int f(x\,|\,\hat\eta)\log(f(x\,|\,\hat\eta)/f(x\,|\,\eta))\mu(dx)$$

is minimized, where μ is a dominating measure for the family Γ. As θ_0 varies, this defines a one parameter family. Since Kullback-Leibler distance is not symmetric, one can create a "forward" or "backward" family using $K(\hat\eta,\eta)$ or $K(\eta,\hat\eta)$, respectively. The corresponding families will be denoted by Γ_θ^F and Γ_θ^B and their log-likelihoods by $L^F(\theta)$ and $L^B(\theta)$.

In section 3 we find the directions of the families at $\hat\eta$. Conditions under which the families are equivalent will be derived. In section 4 we consider two examples.

3. LOCAL EQUIVALENCE OF FAMILIES

The directions of the families can be found using the following lemma.

Lemma: Let $g: \mathrm{R}^k \to \mathrm{R}$ *be three times differentiable with invertible Hessian* $\nabla^2 g$ *and global minimum at* $\hat{\eta} \in \mathrm{R}^k$. *Let* $\theta : \mathrm{R}^k \to \mathrm{R}$ *be continuously differentiable with non-zero gradient on a neighbourhood of* $\hat{\eta}$. *Define a curve c(t) implicitly by*

$$g(c(t)) \equiv \min_{C_t} \; g(\eta)$$

where $C_t = \{\eta : \theta(\eta) = t\}$ *and let* $c(t_0) = \hat{\eta}$. *Then*

$$D(c(t))\,|_{t_0} \equiv \frac{dc(t)}{dt}\,|_{t_0} = \lambda_0 (\nabla^2 g(\hat{\eta}))^{-1} \nabla \vartheta(\hat{\eta})$$

where

$$\lambda_0 = [(\nabla\theta(\hat{\eta}))^t (\nabla^2 g(\hat{\eta}))^{-1} (\nabla\theta(\hat{\eta}))]^{-1}.$$

Proof: First note that c(t) is differentiable by the implicit function theorem (Spivak, 1965, p. 41). Now, since c(t) is defined as a minimum we have (using Lagrange multipliers)

$$\nabla g(c(t)) = \lambda(t) \nabla\theta(c(t)).$$

Differentiation with respect to t gives

$$(\nabla^2 g) D(c(t)) = \lambda(t)(\nabla^2\theta) D(c(t)) + \lambda'(t)\nabla\theta$$

where $\nabla^2 h$ is a matrix with i,j^{th} entry $\partial^2 h / \partial\eta_i \partial\eta_j$ for a function h. Evaluating this expression at t_0 gives the form of $D(c(t))\,|_{t_0}$. (Note that $\lambda(t) = 0$ at $t = t_0$). The constant is determined by differentiating the Lagrange equation with respect to λ then t.

Now let $D_{\hat{\eta}}^{a}$ be the direction vector of a particular family at $\hat{\eta}$ where a = PL,S,F or B to indicate the appropriate family. We have

Theorem: $D_\eta^{PL} = D_\eta^{S} = \alpha(I_{\hat\eta})I_{\hat\eta}^{-1}\,\nabla\theta(\hat\eta)$ *and* $D_\eta^{F} = D_\eta^{B} = \alpha(i_{\hat\eta})i_{\hat\eta}^{-1}\,\nabla\theta(\hat\eta)$ *where*

$$\alpha(A) = [(\nabla\theta(\hat\eta))^t A^{-1}(\nabla\theta(\hat\eta))]^{-1}.$$

Proof: The direction of Γ_θ^S is $(\Gamma^{-1}\nabla\theta)\partial\tau/\partial\theta$ which equals $\alpha(I_{\hat\eta})I_{\hat\eta}^{-1}\nabla\theta(\hat\eta)$ since $\partial\tau/\partial\theta = \alpha(I_{\hat\eta})$. The direction of Γ_θ^{PL} follows from the lemma by taking g to be minus the log-likelihood and assuming the usual regularity conditions. The directions of Γ_θ^F and Γ_θ^B are obtained by noting that to second order terms

$$K(\hat\eta,\eta) = K(\eta,\hat\eta) = \frac{1}{2}(\eta-\hat\eta)^t i_{\hat\eta}(\eta-\hat\eta)$$

(see Kullback (1959, p. 28)). Applying the lemma yields the result.

Therefore Stein's family and the profile likelihood family are locally equivalent as are the two Kullback families. A sufficient condition for $i_{\hat\eta}=I_{\hat\eta}$ is that the model be a member of the exponential family. Hence in this case all four families are locally equivalent. It can also be shown, using Hoeffding's lemma (Efron, 1978) that in the exponential family, the profile family and the forward Kullback family are globally equivalent.

Outside of the exponential family, $I_{\hat\eta}$ and $i_{\hat\eta}$ are in general different; their difference can be expressed as a function of statistical curvature (Efron, 1975 and Skovgard, 1985).

The theorem suggests that inferential techniques based on the local properties of the likelihood function will be similar for all four methods. In particular, note that the second derivative (at $\hat\eta$) of the log-likelihoods is $(D_\eta^a)^t\, I_{\hat\eta}^{-1}(D_\eta^a)$ for a = PL and a = S and is $(D_\eta^a)^t\, i_{\hat\eta}^{-1}(D_\eta^a)$ for a = B and F. Hence $L^{PL}(\theta)$ and $L^S(\theta)$ have the same second derivatives as do $L^F(\theta)$ and $L^B(\theta)$. Agreement of the third derivatives can be shown by a simi-

lar calculation. One application of this result would be the approximation of $L^{PL}(\theta)$ by $L^{S}(\theta)$. This can provide considerable simplification since $L^{S}(\theta)$ requires only the computation of $I_{\hat{\eta}}$ and $\nabla\theta(\hat{\eta})$ while $L^{PL}(\theta)$ requires a restricted maximization at each value of θ. This will be difficult if $\theta(\eta)$ is a complicated function of η. However, the quality of such an approximation is still an open question.

4. EXAMPLES

Example 1.

Let x be bivariate normal with mean η and covariance equal to the identity matrix. Suppose the parameter of interest is $\theta = \eta_1/\eta_2$. Note that θ is constant over rays through the origin in E^2. It is easy to show that $K(\hat{\eta},\eta)$ reduces to 1/2 times the squared Euclidean distance between $\hat{\eta}$ and η so that the forward and backward Kullback-Leibler families and the profile likelihood based family all correspond to the circle passing through the origin and $\hat{\eta}$ (see Figure 1a). The corresponding likelihood functions are plotted in Figure 1b.

Example 2.

This example is motivated by Efron's (1984) use of the least favourable family in computing bootstrap confidence intervals. The data $x_1, x_2, \ldots x_n$ are fixed and the family of rescaled multinomial distributions $M(n,\omega)/n$ is considered. The parameter of interest is a functional $\theta(\omega)$. The natural parameter is $\eta = \log \omega$. A least favourable family is drawn through the m.l.e $\omega_0 = e^{\hat{\eta}} = (1/n, 1/n, \ldots 1/n)$. We illustrate this in Figure 2 for n = 3 with $\theta(\omega) = \bar{x}_\omega = \sum \omega_i x_i$ and $(x_1, x_2, x_3) = (-1,0,1)$. The triangle represents the simplex

$$S_3 = \{\omega \mid \omega_i \geq 0, \sum_1^3 \omega_i = 1\}.$$

The least favourable family and backward Kullback-Leibler family agree while the profile likelihood and forward Kullback-Leibler families coincide. Also shown are the level curves C_{θ_0}.

ACKNOWLEDGEMENTS

We would like to thank the referees for their helpful comments. We would also like to thank Penny Brasher for her help with the illustrations, a previous referee for suggesting the general form of the lemma and Tom DiCiccio for helpful discussions.

REFERENCES

Efron, B. (1975). Defining the curvature of a statistical problem. *Ann. Statist.* **3** , 1189-1242.

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** , 362-376.

Efron, B. (1981). Non-parametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9** , 139-172.

Efron, B. (1984). Better bootstrap confidence intervals. Stanford University Technical Report 226.

Kalbfleisch, J.D. and Sprott D.A. (1970). Applications of likelihood methods to models involving large numbers of parameters. *J.R. Statist. Soc.*, **32** , 175-208.

Kullback, S. (1959). *Information theory and statistics.* John Wiley and Sons, Inc., N.Y., N.Y.

Skovgard, I. M. (1985). A second order investigation of asymptotic ancillarity. *Ann. Statist.*, **13** , 534-551.

Spivak, M. (1965). *Calculus on manifolds.* W.A. Benjamin Inc., N.Y., N.Y.

Stein, C. (1956). Efficient non-parametric testing and estimation. *Proceedings of the 3rd Berkeley Symposium.* 187-196.

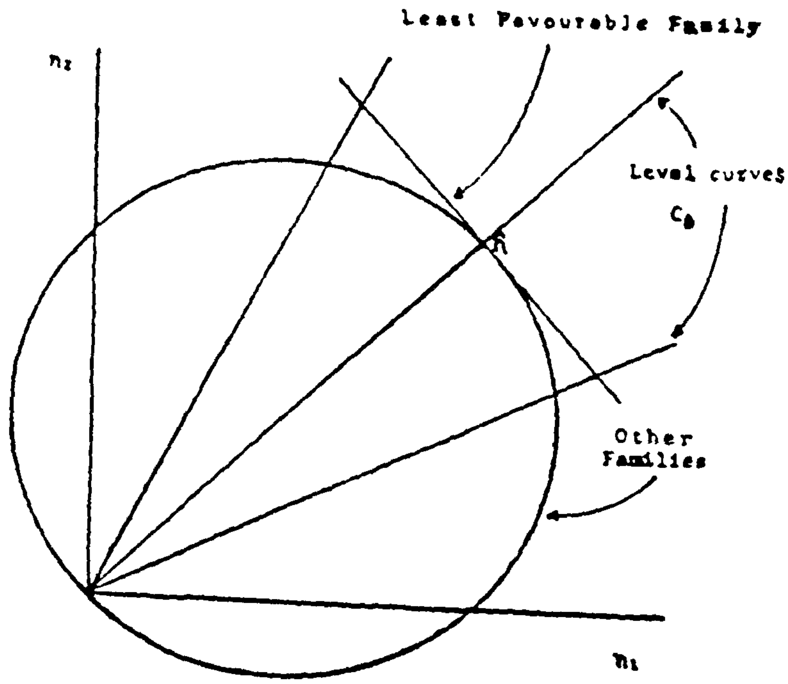


FIG 1a Families for $x \sim N(\eta_1, \eta_2, I)$ $\theta = \eta_1 / \eta_2$

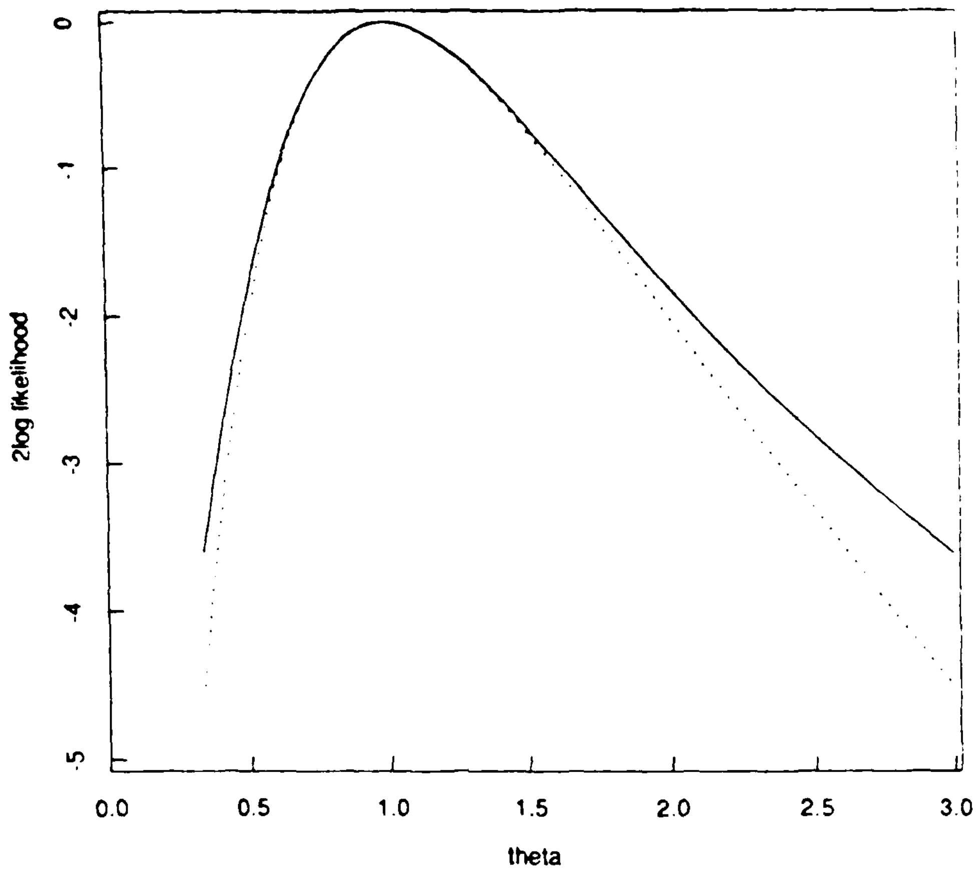


FIG 1b Twice log-likelihood for least favourable family (dotted line) and other families (solid line) corresponding to problem of Figure 1.

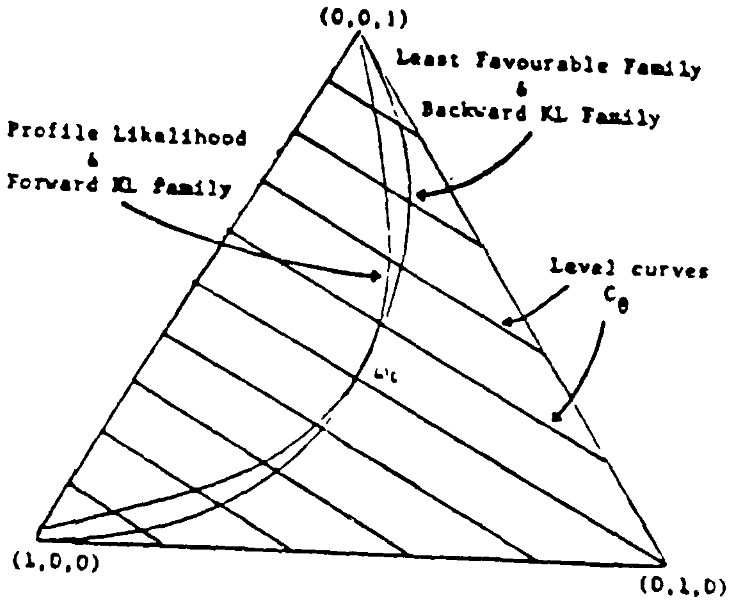


FIG 2 Families for multinomial, $\theta = \bar{x}$

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 404	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Note On Profile Likelihood, Least Favourable Families And Kullback-Leibler Distance		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert Tibshirani and Larry Wasserman		8. CONTRACT OR GRANT NUMBER(s) N00014-86-K-0156
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE April 19, 1988
		13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Nuisance parameters, likelihood, exponential families.		

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several methods exist for reducing higher dimensional problems to a single parameter. These include the profile likelihood, least favourable families and methods based on the Kullback-Leibler distance function. We demonstrate that at least in a neighborhood of the maximum likelihood estimate, these are equivalent for the exponential family.

END
DATE
FILMED
DTIC
6-88